ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Laboratory for Computer communications and Applications 1

LCA1

Erman Ayday, Jean Louis Raisaro, Mathias Humbert and Jean-Pierre Hubaux

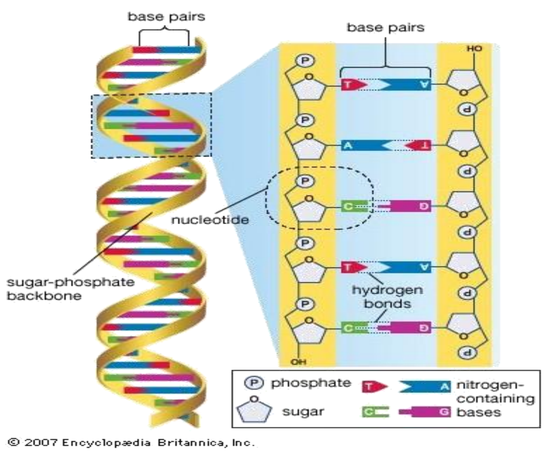
Towards Quantifying and Preventing the Leakage of Genomic Data Using Privacy-Enhancing Technologies

1. Introduction and Goals

A full genome sequence not only uniquely identifies each one of us; it also contains information about our ethnic heritage, disease predispositions, and many other phenotypic traits.

Threats due to genomic data leakage:

- Revelation of predisposition to diseases, ethnicity, paternity, filiation, etc.
- Genetic discrimination.
- Denial of access to health insurance, mortgage, education, and employment.

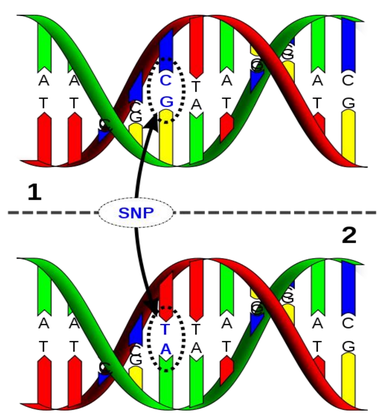


Goals:

- Protect the privacy of users' genomic data.
- Protect the privacy of medical unit's confidential data.
- Allow specialists to access only to the genomic data they need (or they are authorized for).
- Keep the access time to a single patient's genomic data to a few seconds.

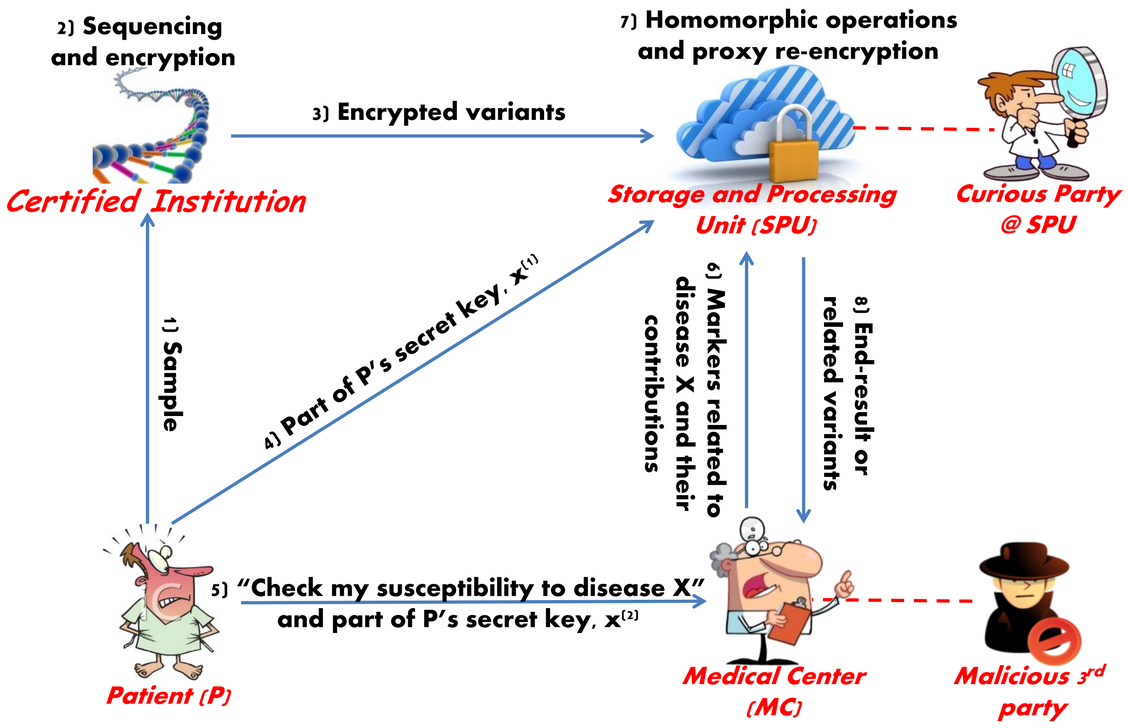
2. Genomic Background

The human genome has approximately 3 billion letters.



- Single Nucleotide Polymorphisms (SNPs): DNA variations, occurring when a single nucleotide differs between members of the same species.
- Potential nucleotides for a SNP position are called alleles.
- A disease susceptibility test is done by analyzing particular SNPs.
- Each SNP contributes to the disease susceptibility in a different amount.
- 40 million approved SNPs in the human population.
- Each patient carries around 4 million SNPs out of 40 million – real SNPs of the patient.
- SNPs are correlated to each other depending on the Linkage Disequilibrium (LD) property.

3. Proposed Solution

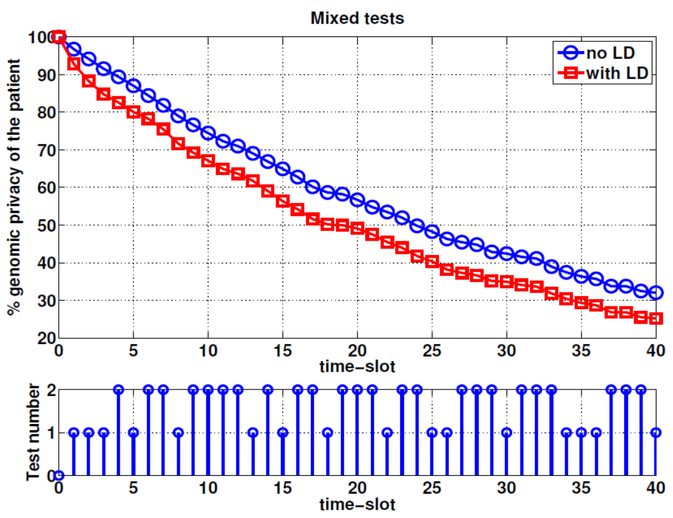


- Store patients' genomic data encrypted by their public keys at the Storage and Processing Unit (SPU).
- Process the encrypted genomic data for medical tests and personalized medicine methods using homomorphic re-encryption and proxy encryption.

4. Threat Model

- A **curious party** at the Storage and Processing Unit (SPU), who tries to infer the genomic sequence of a patient from his stored data.
- A **malicious medical center**, who can be considered either as an attacker that hacks into the medical center's system or a disgruntled employee, who happens to access the medical center's database.

5. Quantification of Genomic Privacy



ASYMMETRIC ENTROPY

$$h(p_i) = \frac{p_i(1-p_i)}{(-2w+1)p_i + w^2}$$

At the Medical Center:

- From the results of genetic tests, the MC can learn more information than it is authorized by using:
  - Characteristics of the exposed SNPs (LD).
  - Disease markers and their contributions.
- Quantifying genomic privacy by computing the decrease in privacy of a patient as a result of the number of tests he undergoes.
  - **Test 1**: MC obtains a subset of P's SNPS.
  - **Test 2**: MC obtains the end-result of a genetic test.
  - Use of asymmetric entropy as genomic privacy metric.
  - p_i is the probability of correctly inferring the content of a SNP and w is the point at which the entropy is maximum.

6. Implementation and Computational Complexity

- Intel Core i7-2620M CPU with 2.70 GHz processor.
- Size of the security parameter: 4096 bits.
- Real SNP profiles from 1000 Genomes Project.
- Java programming language.

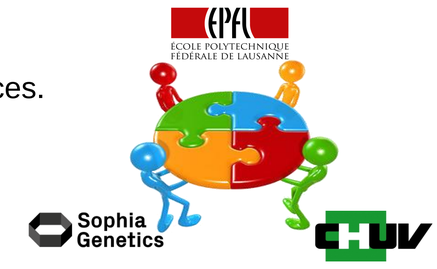
@CI	@ SPU			@MC
Paillier Encryption	***Homomorphic operations***	***Proxy Re-encryption***	***Storage***	***Paillier decryption***
380 ms. per variant	**100 sec. (10 variants)**	**387 ms.**	**4,1 GB per patient**	**1,4 sec.**

7. Discussion and Future Work

- **Future work**: develop new techniques to prevent the leakage of genomic data.
- **Encourage the use of genomic data, by the individual and by the medical center, and accelerate the move of genomics into clinical practice.**

Collaboration:

- EPFL - School of Computer and Communication Sciences.
- EPFL - School of Life Sciences.
- University Hospital of Lausanne (CHUV).
- Sophiagenetics.com.



Web link: http://lca.epfl.ch/projects/genomic-privacy/